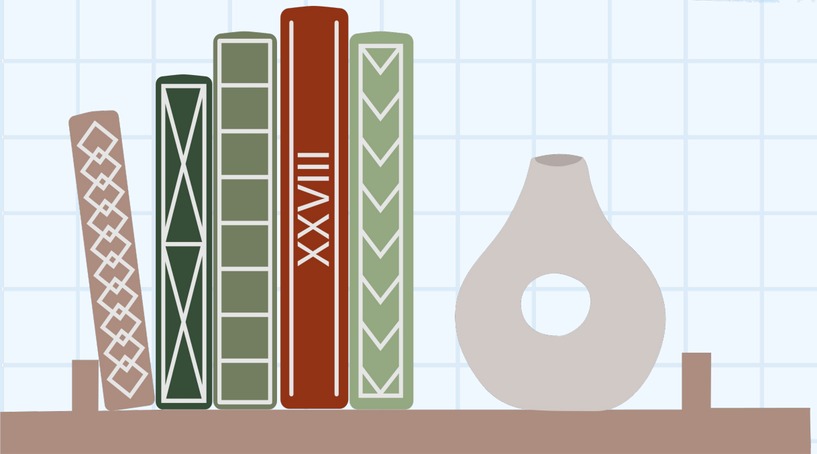
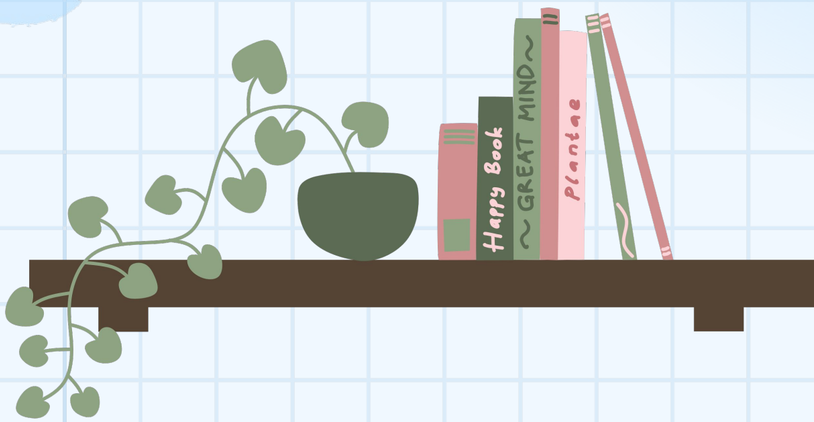
Powered by
Futurense

BS./BSC.IN

Applied AI and Data Science

Basics of Data Analytics





Powered by Futurense

Let's dive into and learn:

1. Extract, Transform, Load
2. Data Pipeline

Data Pipelines

Powered by

Futurense

- Data Processing pipeline is often called ETL pipeline
- **E**xtract, **T**ransform, **L**oad
- Process by which raw data is converted into analysis ready data

ETL

- Gathering raw data
- Extracting information needed for reporting and analysis
- Cleaning, organizing, standardizing, formatting data into usable format
- Loading the data into a repository

Extract

- Step through which data from the originating location is collected for transformation
- Can be processed in batches
    - Data is collected and moved from source to destination in chunks at regular intervals
    - Data can also be pulled from source to destination continually in real-time with some processing in the way

Transform

- Transform is the process by which raw data is converted into a format that can be readily used for analysis
- E.g.
  - Removing duplicates
  - Making date formats consistent
  - Changing formats for ease of storage
  - Creating key relationships across tables

Load

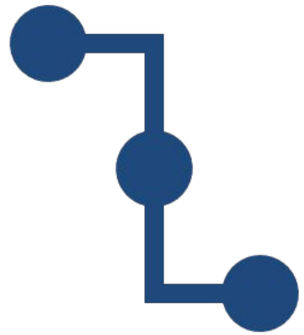
- Process by which cleaned data is transported to the destination repository for analysis
- Entire data needed for analysis could be loaded at once – initial loading
- Sections of the data could be loaded periodically – incremental loading
- Erasing the previously loaded contents and loading fresh data – full refresh

Caution while loading

- Load verification includes
  - Checking for missing or null values
  - Server performance
  - Monitoring load failures

Data Pipeline

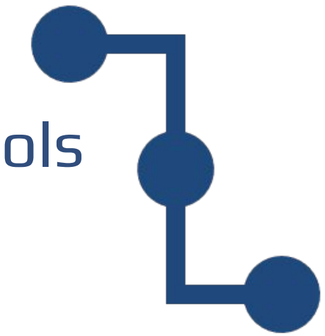
- Data Pipeline is a broad term
- Refers to the entire journey of transferring data from one system to another
- ETL is a part of the data pipeline
- Data Pipeline can be for multiple types of processing – batch or streaming



Powered by Futurense

Data Pipeline

- Data Pipeline is a high-performance system
- Supports both batch queries and smaller interactive queries
- May load data into a data lake
- May load data onto other applications or visualization tools
- E.g. Apache Beam , Data Flow



Recap

Powered by

Futurense

Understanding the process of

ETL

Data pipeline

Powered by
Futurense

Thank you